

COMP309: Machine Learning Tools and Techniques

Exploratory Data Analysis (EDA)

- Tutorial 5
- week 5
- Baligh Al-Helali

What questions need to be asked about the data before going ahead with processing and modelling?

1. What is the purpose of the data?
2. Which variables are predictive features and what is the target one?
3. What is the size/scale of the data?
4. What are the types of the variables (features and target)
5. Are the data clean?
6. Are the data balanced?
7. The distribution of each variable?
8. How the relationships between the variables look like?
9. Are there any correlations between variables?
10. Are there any missing values?
11. Which features might be more important?
12. ...

Why such questions help for processing and modelling?

❑ Understand the data helps:

1. Easier to spot mistakes
2. Easier to set hypotheses
3. Allows you to feel your data
4. Get an idea of how/what the experiment should be conducted
5. ...

❑ Examples of important decisions based on EDA:

- Type of target → classification or regression
- Data balance → performance measure (accuracy, weighted, F1-score)

How to answer such questions?

❑ Some questions require reading the documents of the data creators (e.g. Qs 1 and 2)

- Reading the data info and details from the providing websites (e.g. UCI and OpenML)
- Visiting the creators website or reading the paper(s) that published the data (if any)
- Contacting the data creators (e.g. email)

❑ Some questions require extracting statistics and visualisation such as Qs 3-11

- Statistics: count, mean, median, quartiles, variance, ...
- Visualisation: histograms, scatterplot, boxplot, heatmap, ...

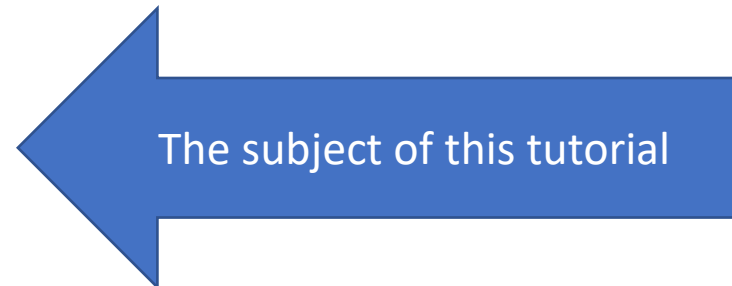
❑ How to get the stats and Vizs:

- From the data sources (Not always reliable)

- Using EDA tools:

- Orange
- Python: Pandas, Matplotlib, Seaborn

- ...



We are considering the popular data set “iris”

-The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.^[1](Wikipedia)



❑ UCI Machine Learning Repository

Iris Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Famous database; from Fisher, 1936

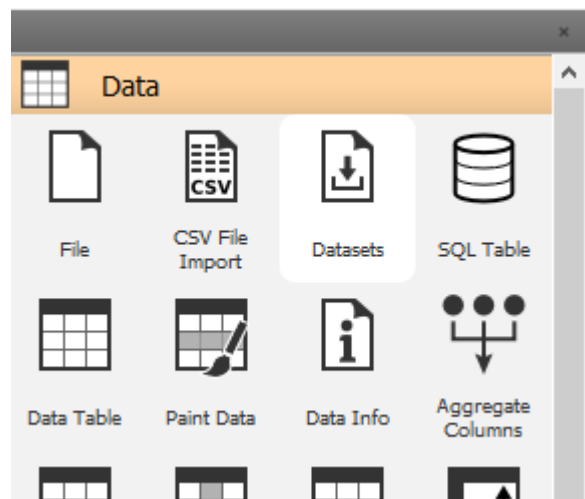


Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	4128602

[1]. R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". *Annals of Eugenics*. 7 (2): 179–188. [doi:10.1111/j.1469-1809.1936.tb02137.x](https://doi.org/10.1111/j.1469-1809.1936.tb02137.x). [hdl:2440/15227](https://hdl.handle.net/2440/15227).

Let's start with Orange

- Load the data set

A screenshot of the 'Datasets' widget in the Orange3 software. The widget has a title bar 'Datasets' and a search bar. Below the search bar is a table listing various datasets. The 'Iris' dataset is selected and highlighted. Below the table is a description of the selected dataset, including its origin and a reference to the original paper.

Title	Size	Instances	Variables	Target	Tags
Iris	4.5 KB	150	5	C categorical	biology
Breast Cancer and Docetaxel Treatment	1.8 MB	24	9486	C categorical	biology
Smoking effect on B lymphocytes	1.8 MB	79	3000	C categorical	genomics
Bone marrow mononuclear cells with AML	582.0 KB	96	1000	C categorical	genomics
HDI	65.1 KB	188	66	N numeric	economy, geo
TKI resistance	1.2 MB	280	467	C categorical	spectral
Abalone	187.5 KB	4177	8	N numeric	biology
Adult	4.1 MB	32561	15	C categorical	economy
Roman Amphorae	23.7 KB	164	16	C categorical	archaeology, image analytics
Attrition - Predict	838 bytes	3	18	C categorical	economy, synthetic, education
Attrition - Train	182.2 KB	1470	18	C categorical	economy, synthetic
Auto MPG	17.3 KB	398	9	N numeric	

Description

Iris (1936), from [UCI ML Repository](#)

The Iris flower data set or Fisher's Iris data set was introduced by the British statistician and biologist Ronald Fisher in his 1936 paper as an example of linear discriminant analysis. The data on length and width of petal and sepal leaves was actually collected by American botanist Edgar Anderson to quantify the morphologic variation of Iris flowers of three related species.

See Also

[Scatter Plots: the Tour](#).

[All I See is Silhouette](#).

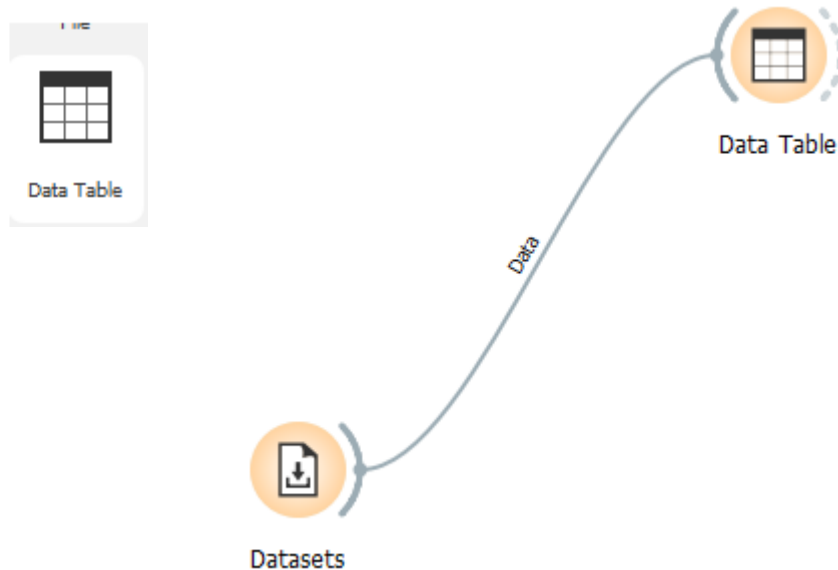
References

R. A. Fisher (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2):179-188

□ What to notice?

- Type of target is categorical → classification
- Data size is small → might need cross validation

Orange EDA: A first look



Data Table

Info
150 instances (no missing data)
4 features
Target with 3 values
No meta attributes

Variables
☒ Show variable labels (if present)
☒ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order
☒ Send Automatically

	iris	sepal length	sepal width	petal length	petal width
40	Iris-setosa	5.1	3.4	1.5	0.2
41	Iris-setosa	5.0	3.5	1.3	0.3
42	Iris-setosa	4.5	2.3	1.3	0.3
43	Iris-setosa	4.4	3.2	1.3	0.2
44	Iris-setosa	5.0	3.5	1.6	0.6
45	Iris-setosa	5.1	3.8	1.9	0.4
46	Iris-setosa	4.8	3.0	1.4	0.3
47	Iris-setosa	5.1	3.8	1.6	0.2
48	Iris-setosa	4.6	3.2	1.4	0.2
49	Iris-setosa	5.3	3.7	1.5	0.2
50	Iris-setosa	5.0	3.3	1.4	0.2
51	Iris-versicolor	7.0	3.2	4.7	1.4
52	Iris-versicolor	6.4	3.2	4.5	1.5
53	Iris-versicolor	6.9	3.1	4.9	1.5
54	Iris-versicolor	5.5	2.3	4.0	1.3
55	Iris-versicolor	6.5	2.8	4.6	1.5
56	Iris-versicolor	5.7	2.8	4.5	1.3
57	Iris-versicolor	6.3	3.3	4.7	1.6
58	Iris-versicolor	4.9	2.4	3.3	1.0
59	Iris-versicolor	6.6	2.9	4.6	1.3

? | 150 | 150 | 150

What to notice?

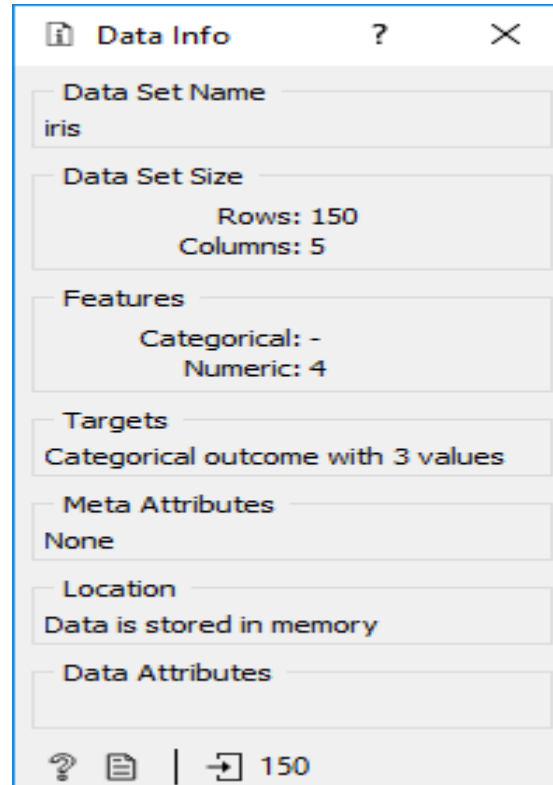
- Variable scale is different (e.g. sepal length is the widest and petal width is the least)
- ➔ might need normalization

150 | 150

Selected Data: iris: 150 instances, 5 variables
Features: 4 numeric (no missing values)
Target: categorical

Data: iris: 150 instances, 6 variables
Features: 4 numeric (no missing values)
Target: categorical
Metas: categorical

Orange EDA: A first look

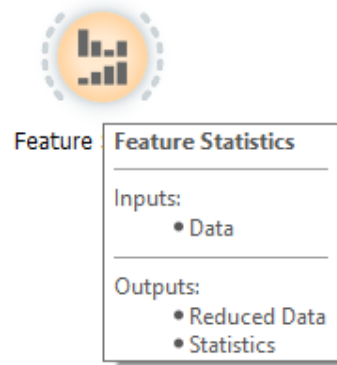


❑ What to notice?

- Type of target is categorical → classification
- Data size is small → might need cross validation

Orange EDA:

What are the stats of the variables?



❑ What to notice?

- Distributions of variables over classes
- Centre, spread, no missing values of variables
- Classes balance

Orange EDA:

What are the most important features?



Rank

Scoring Methods

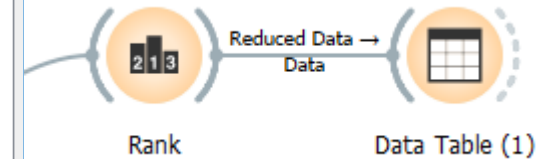
- ☒ Information Gain
- ☒ Information Gain Ratio
- ☒ Gini Decrease
- ☒ ANOVA
- ☒ χ^2
- ☒ ReliefF
- ☒ FCBF

Select Attributes

- ☐ None
- ☐ All
- ☐ Manual
- ☒ Best ranked: 5

☒ Send Automatically

	#	Info. gain	Gain ratio	Gini	ANOVA	χ^2	ReliefF	FCBF
N petal length		1.086	0.544	0.423	1179.034	98.946	0.368	1.542
N petal width		1.059	0.532	0.407	959.324	94.162	0.370	1.451
N sepal length		0.624	0.313	0.247	119.265	79.243	0.155	0.000
N sepal width		0.361	0.183	0.154	47.364	50.082	0.118	0.255



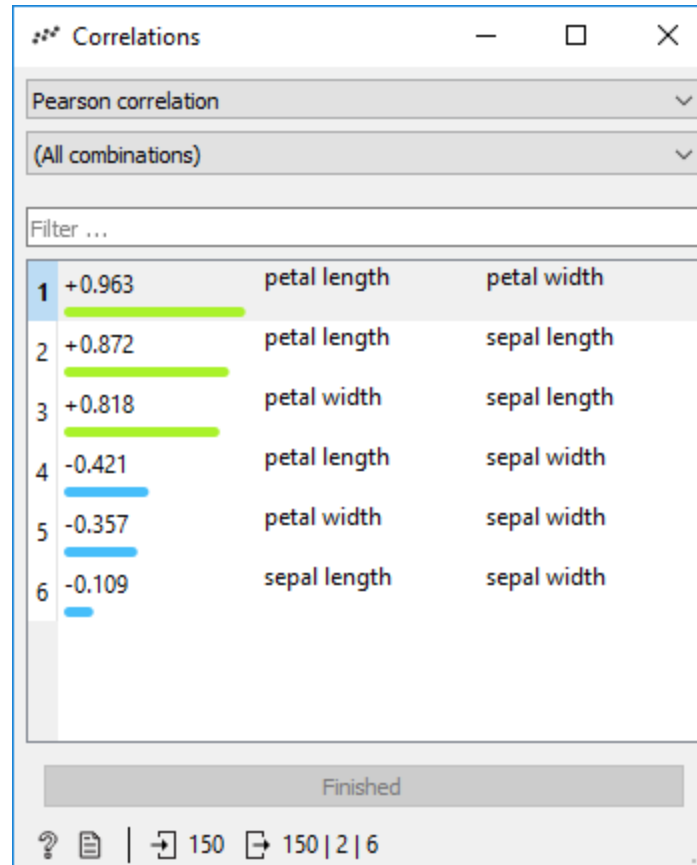
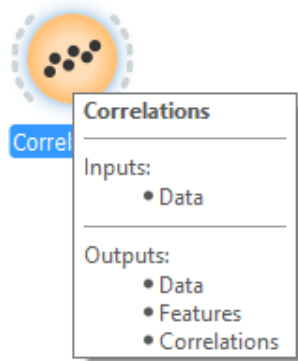
What to notice?

- Petal length seems the most important and sepal width is the least

➔ Feature selection

Orange EDA:

What are the most important features?

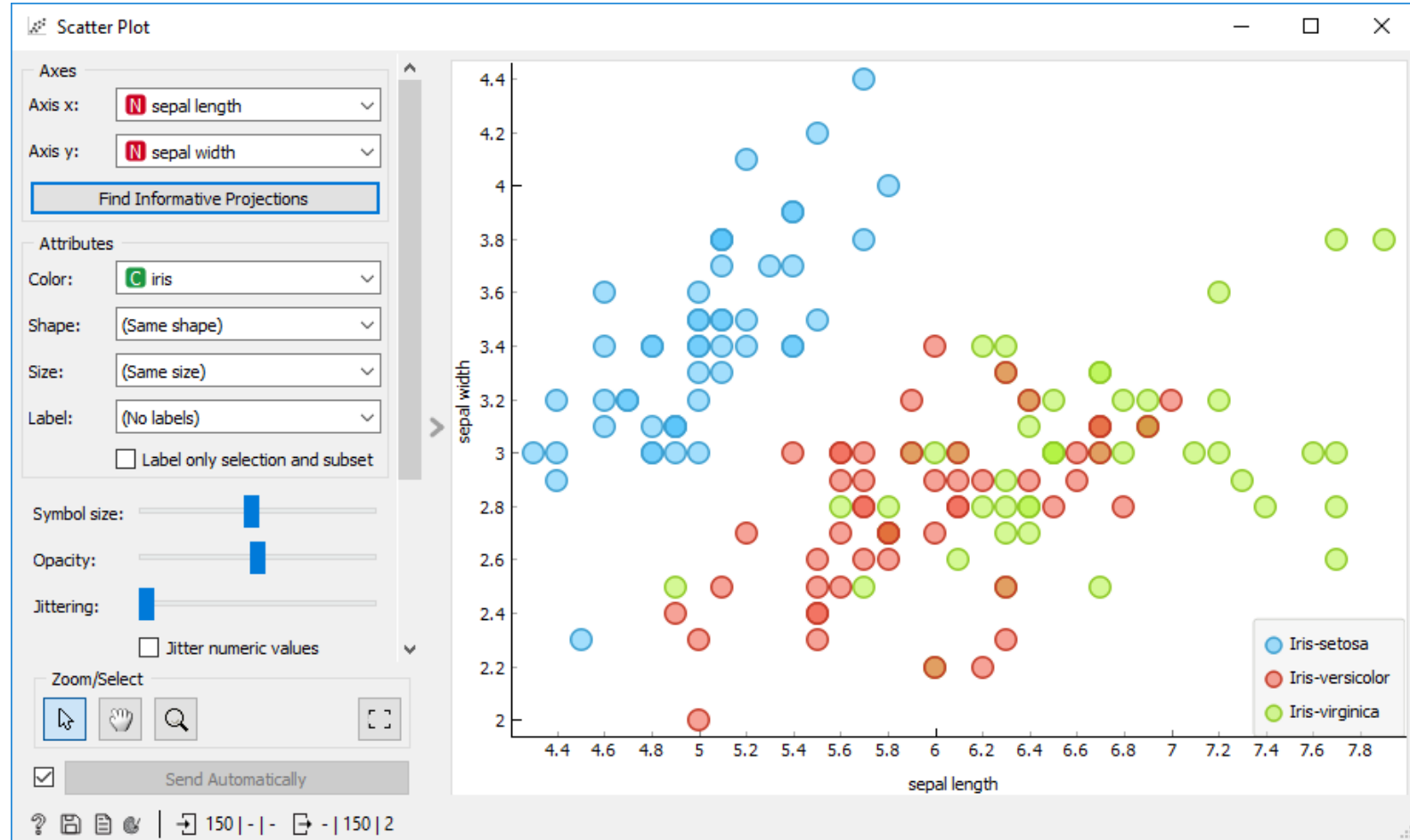


❑ What to notice?

- [-1, 1], negative/positive, strong/weak...
 - Petal length and petal width look strongly correlated
- ➔ Feature selection

Orange EDA:

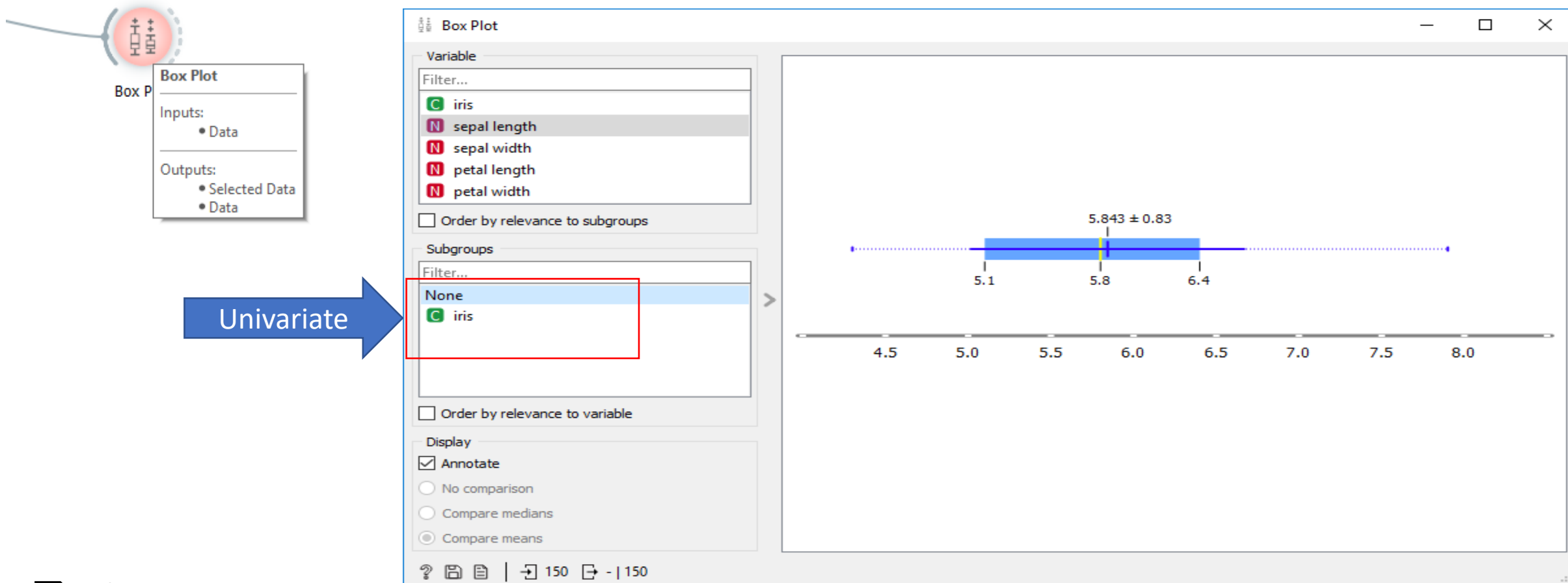
What is the relationship between two variables (e.g. the sepal length and width) per/regardless class?



- Change variables
- What to notice?
- Compare with the correlation shown previously

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed?

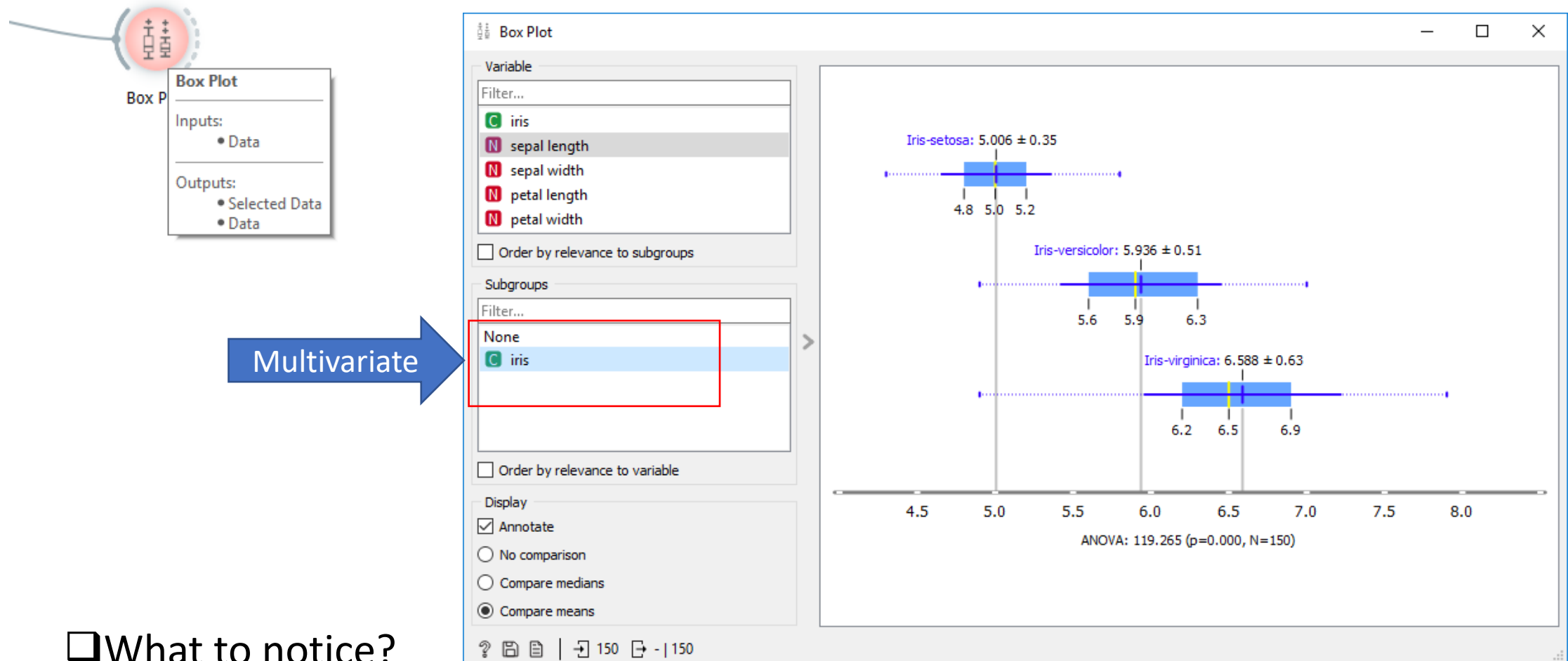


❑ What to notice?

- Graphical presentation for the stats

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed per target class (iris species)?



❑ What to notice?

- Graphical presentation for the stats per class
- Small sepal length → Iris-setosa class

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed per target class (iris species)?



Violin Pl

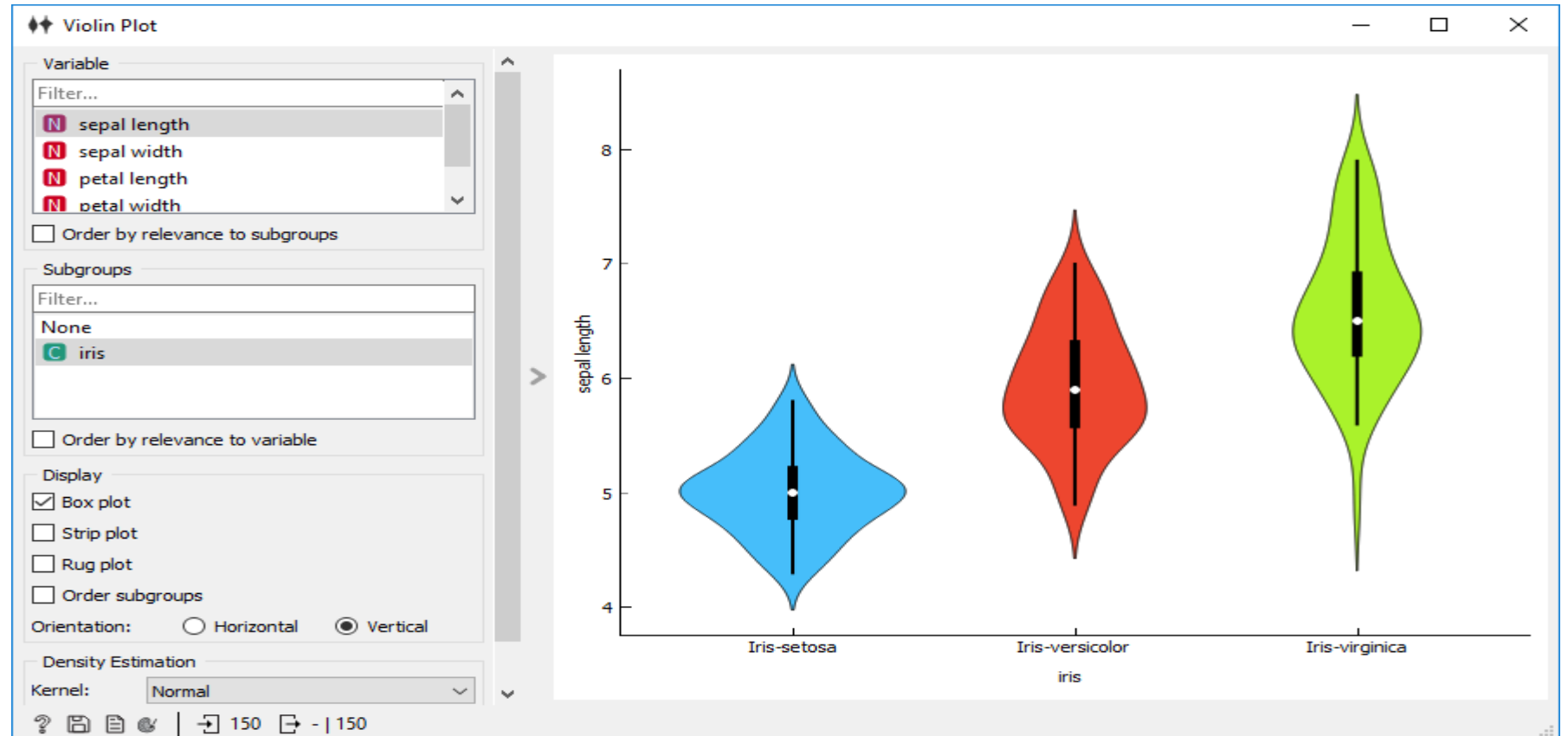
Violin Plot

Inputs:

- Data

Outputs:

- Selected Data
- Data

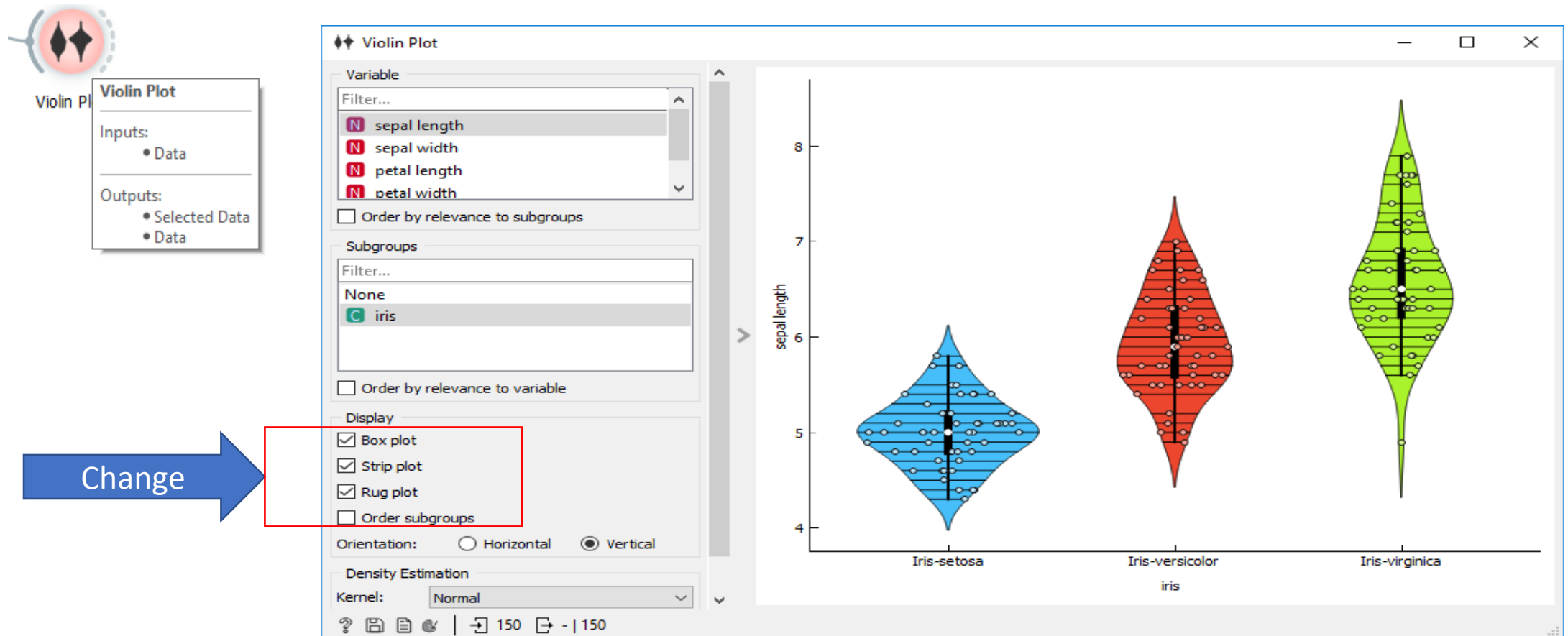


❑ What to notice?

- Similar to box plot but the density/frequency of the samples for variable values is visualized

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed per target class (iris species)?

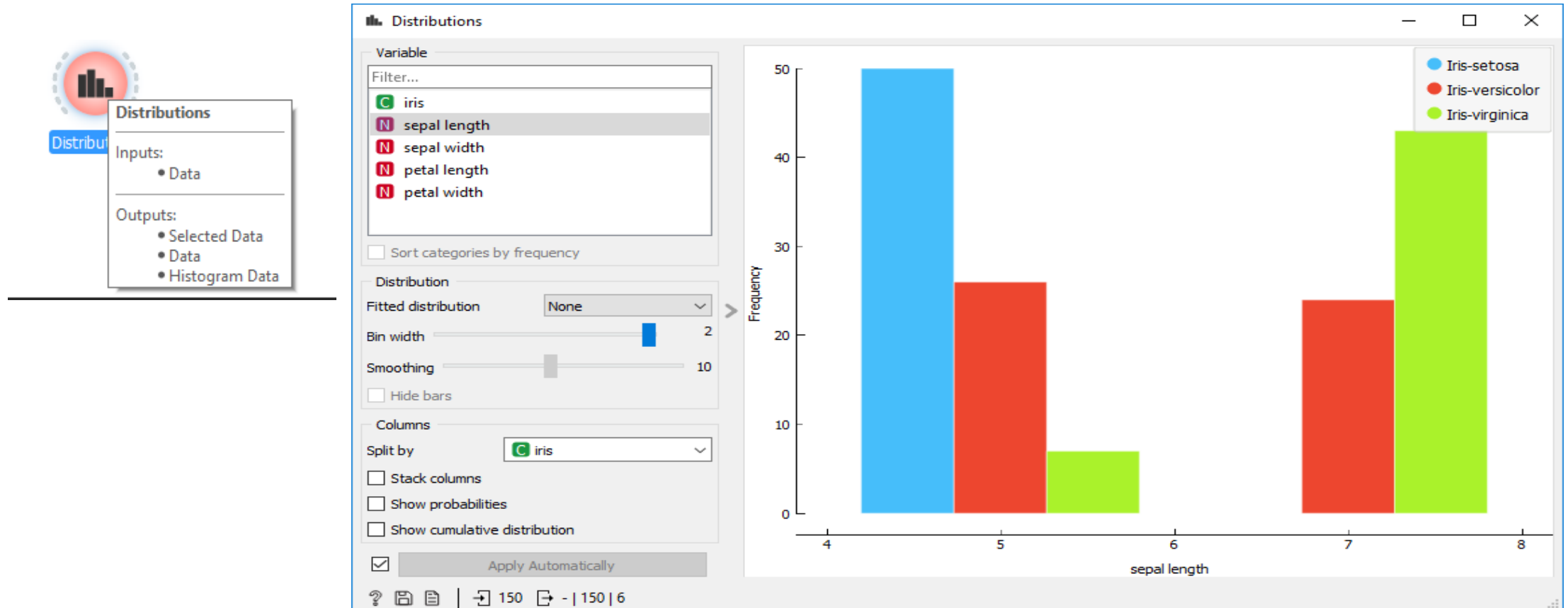


❑ What to notice?

- Show the points for clearer visualization

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed per target class (iris species)?

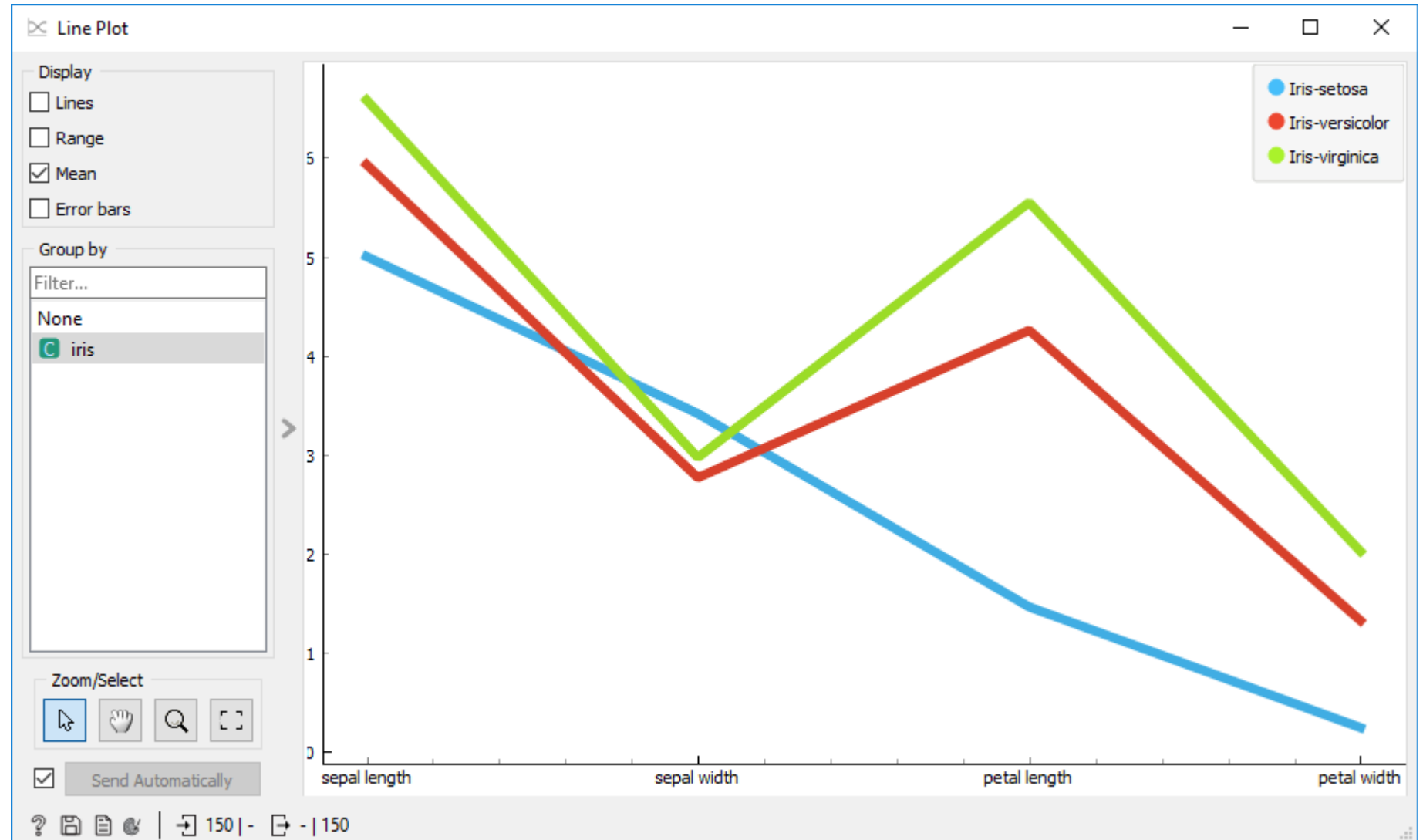


❑ What to notice?

- Shorter sepal → Iris-setosa
- Longer sepal → more likely Iris-virginica

Orange EDA:

How the values of input variables are distributed per target class?

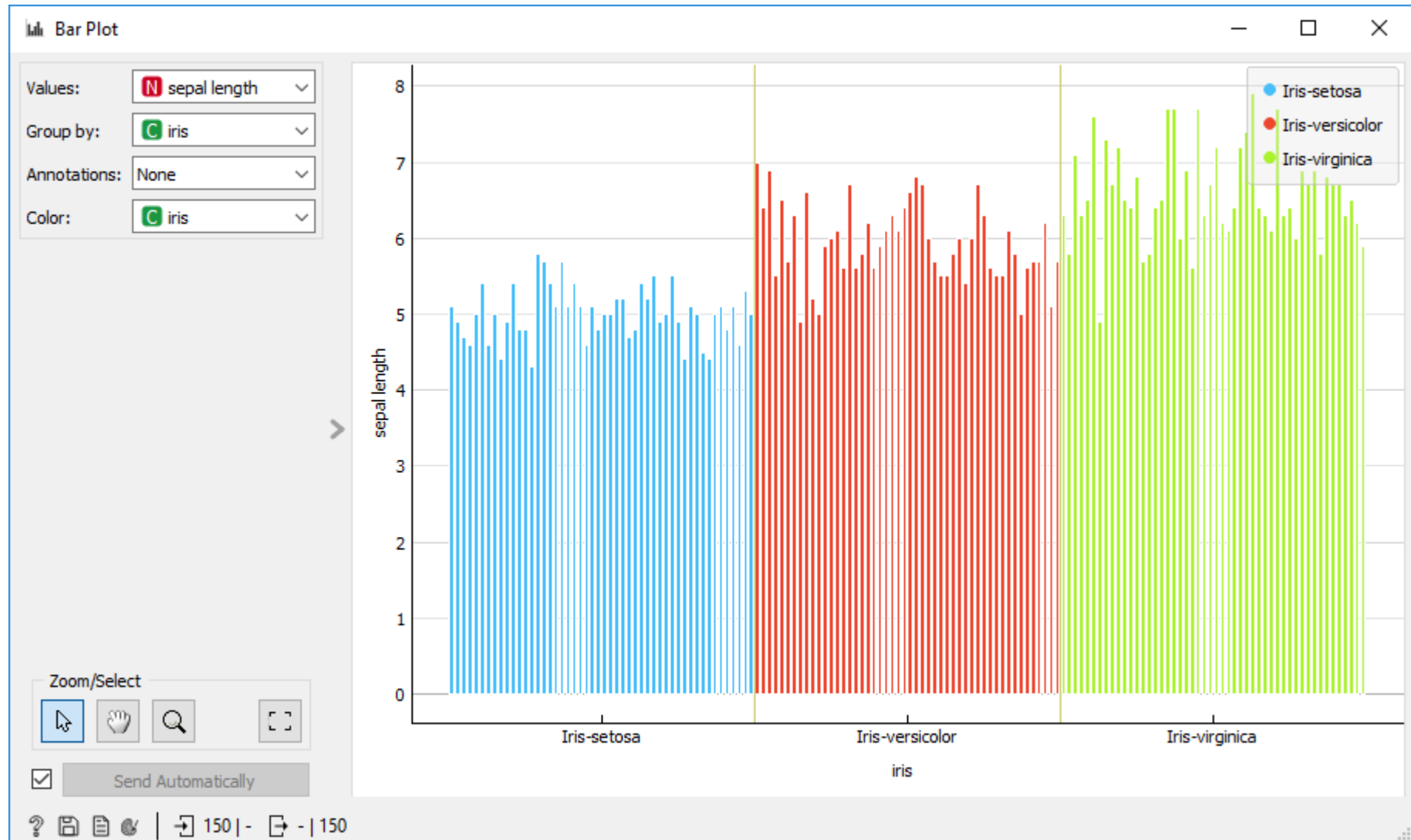


Orange EDA:

How the values of input variables are distributed w.r.t. another variable?



Bar Plot



Orange EDA

- Why there are different ways to answer the same question?